

Future Computer & Programming Trends

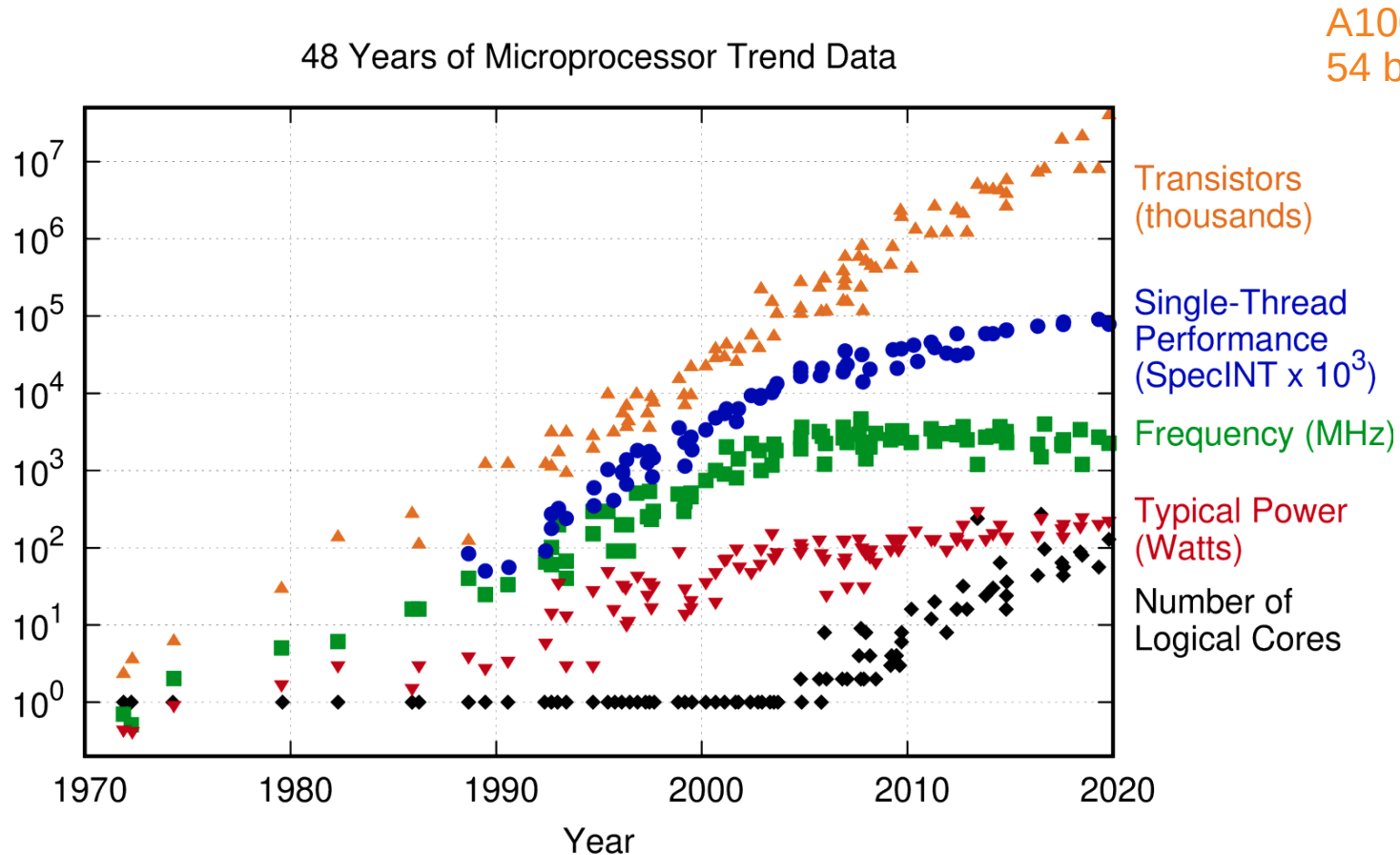
Axel Huebl

Lawrence Berkeley National Laboratory, U.S.

Snowmass Community Planning Meeting

Session 64: Computing Needs of the Accelerator Frontier – Oct 6th, 2020

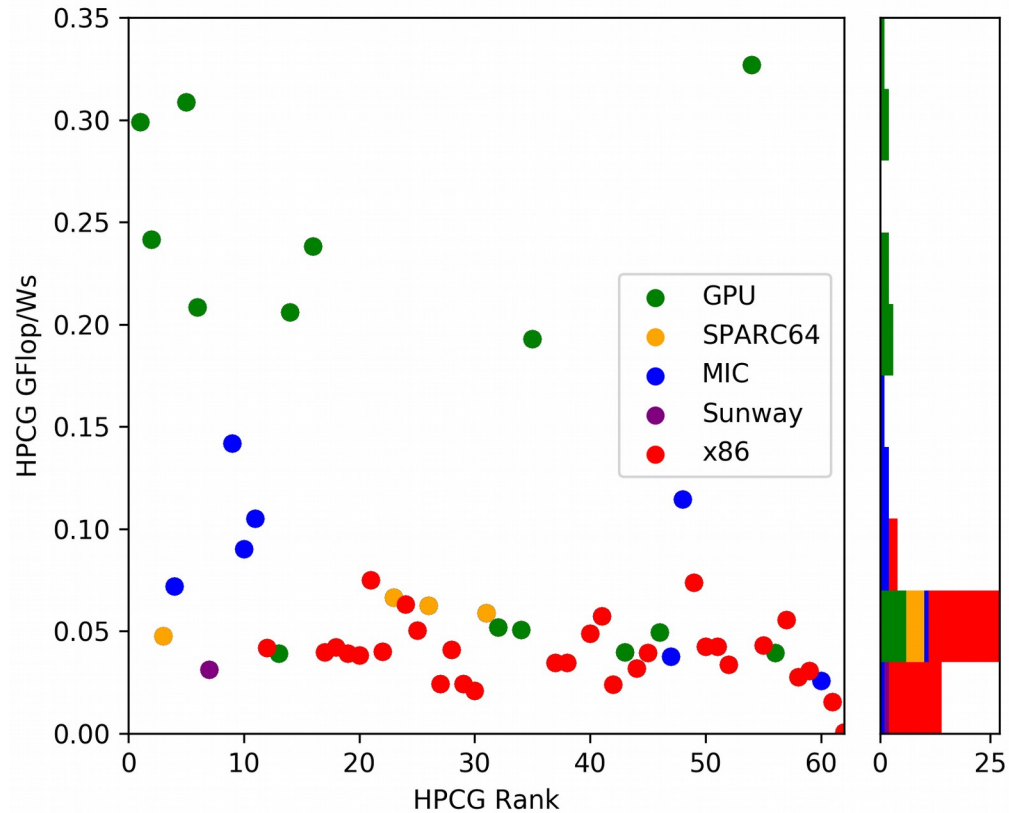
48 Years of Microprocessor Trend Data



A100 GPU:
54 billion transistors

Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2019 by K. Rupp

Power Consumption: HPCG Benchmark



A. Huebl "The Green HPCG List," (11/2018)

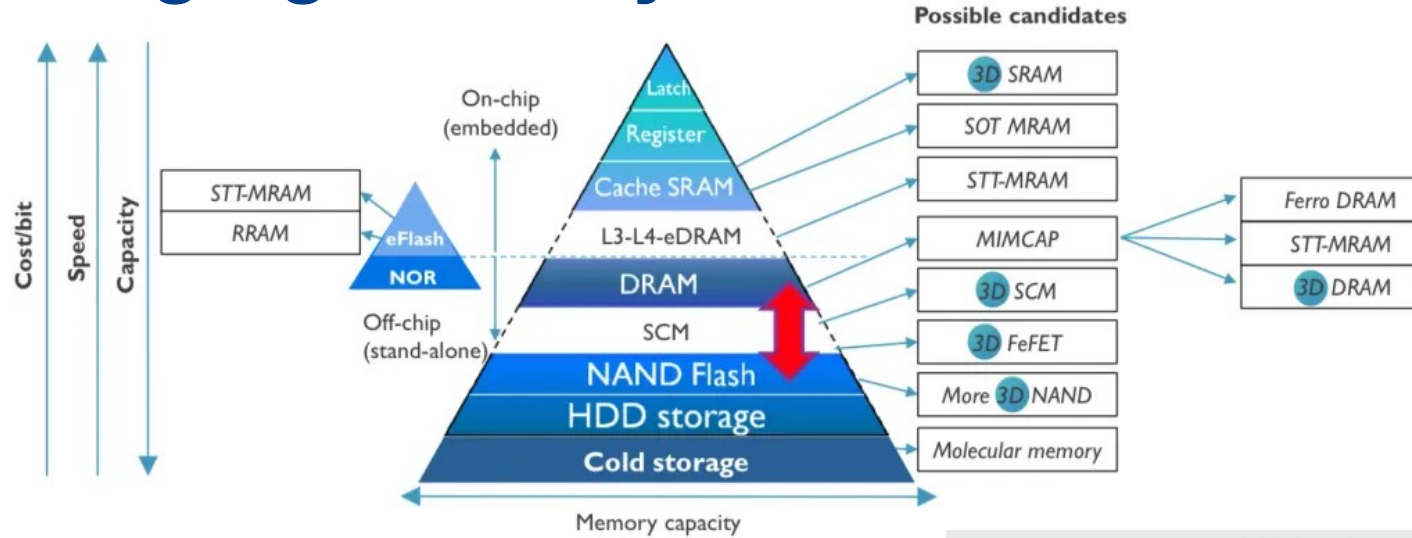
https://plasma.ninja/blog/hpc/manycore/top500/computing/hardware/energy/efficiency/2018/11/18/HPCG_Green.html

Growing Divide: Data & Compute



John Backus, 1977 ACM Turing Award:
„much of that traffic concerns not significant
data itself, but where to find it“

Emerging Memory to Interconnect

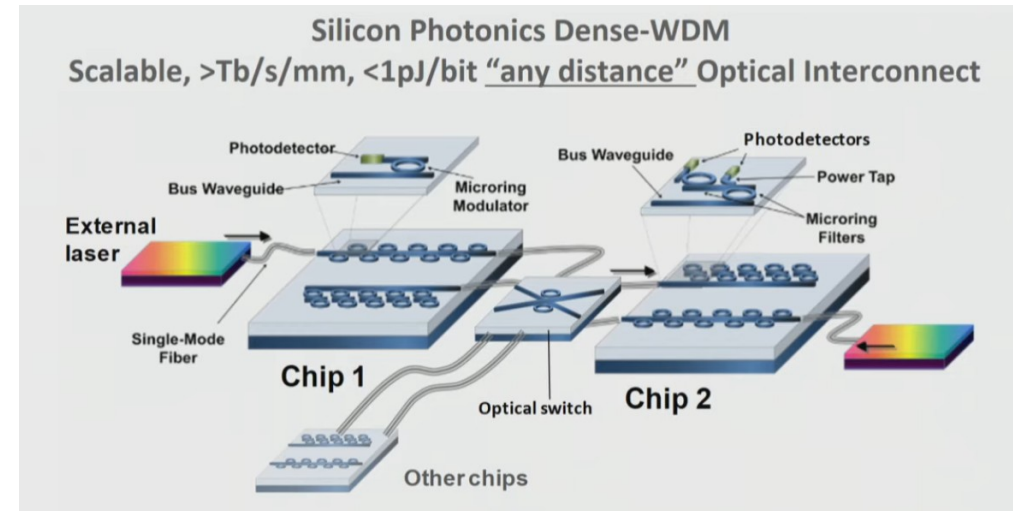


L1 cache reference	0.5 ns	
Branch mispredict	5 ns	
L2 cache reference	7 ns	
Mutex lock/unlock	25 ns	
Main memory reference	100 ns	
Compress 1K bytes with Zippy	3,000 ns	= 3 μ s
Send 2K bytes over 1 Gbps network	20,000 ns	= 20 μ s
SSD random read	150,000 ns	= 150 μ s
Read 1 MB sequentially from memory	250,000 ns	= 250 μ s
Round trip within same datacenter	500,000 ns	= 0.5 ms
Read 1 MB sequentially from SSD*	1,000,000 ns	= 1 ms
Disk seek	10,000,000 ns	= 10 ms
Read 1 MB sequentially from disk	20,000,000 ns	= 20 ms
Send packet CA->Netherlands->CA	150,000,000 ns	= 150 ms

IMEC: <https://semiengineering.com/a-new-memory-contender/>

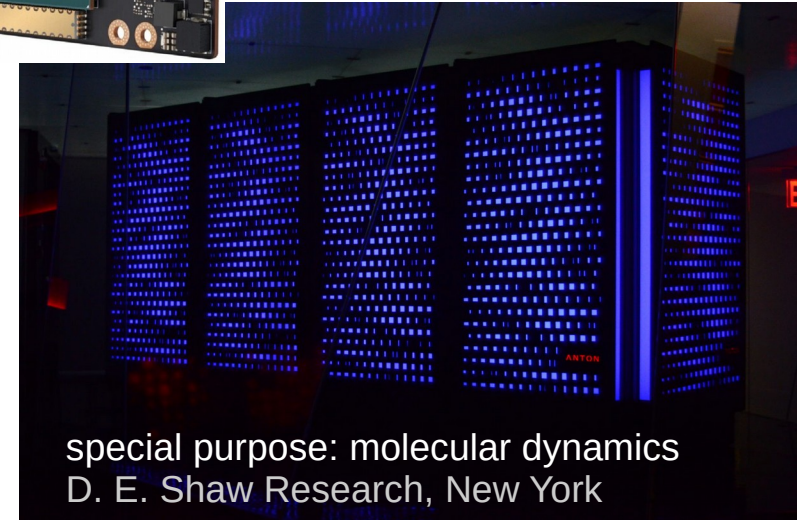
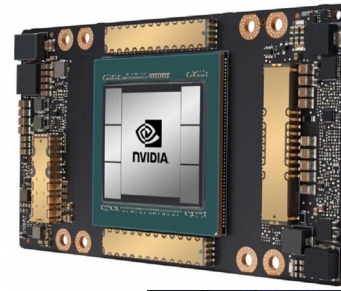
K. Bergmann "Flexibly Scalable High Performance Architectures with Embedded Photonics" PASC Keynote (2019)

P. Norvig <http://norvig.com/21-days.html#answers> (2012) https://colin-scott.github.io/personal_website/research/interactive_latency.html



Today and Near-Term: ~5 years

- **Parallelism:** nodes → devices → rings/SMs → cores → (hyper)threads → SIMD-steps... *on* local → shared → cached → global → remote **memory**
- **Hardware Specialization**
 - SIMD: vector to matrix-processing units (tensor cores)
 - whole device:
 - RISC:
 - GPUs: massive parallelism
 - ARM / RISC-V / NEC
 - FPGAs, DSPs, ...
 - ASICs; ANTON2 (2008 → 2014)
- **Algorithmic Specialization**
 - multi-level parallelism; in situ algorithms



special purpose: molecular dynamics
D. E. Shaw Research, New York

Mid- to Long term: >5-10 years

- **Further Specialization**

- Programmable FPGAs from *high-level languages* (“HLS”)
- on-**socket** integration of “<5 year” hardware
- *workload-specific* memory & system *designs*

- **Programming Models**

- *Parallelism* will only rise: width and depth
- C++23 et al.: unification of many of today’s capabilities
- Emerging *new paradigms* likely – for *abstract compute*

- **Potentially non-von Neumann *components***

- First signs: FPGAs, DSPs, memory-driven algorithms, neuromorphic chips, ...

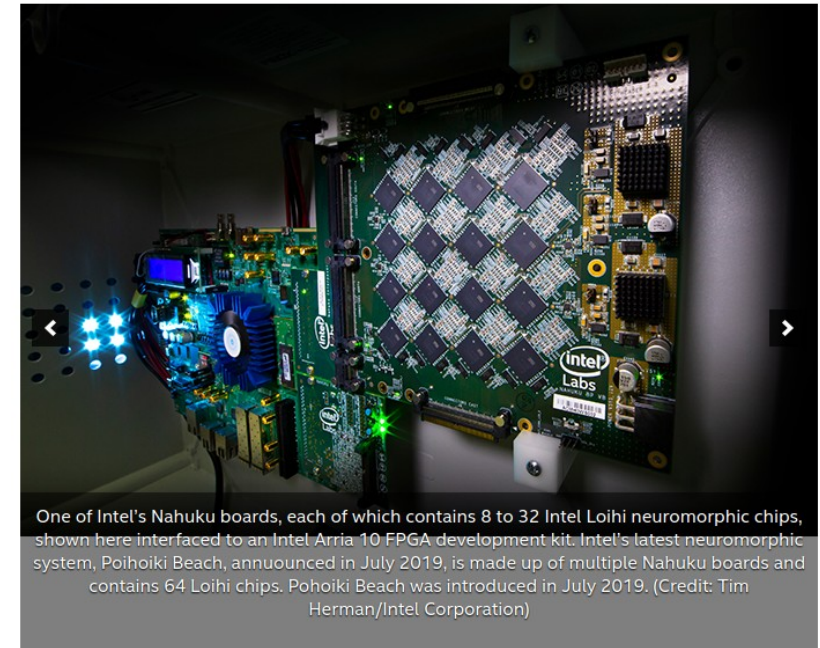
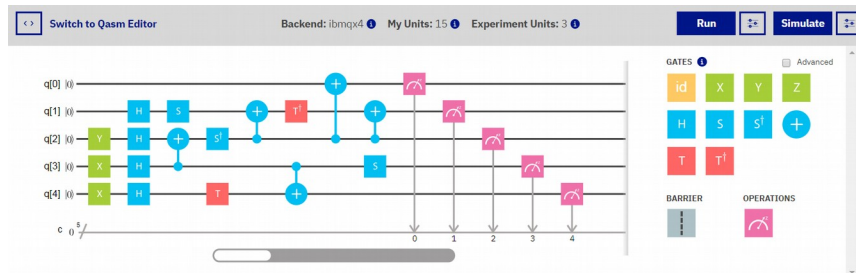
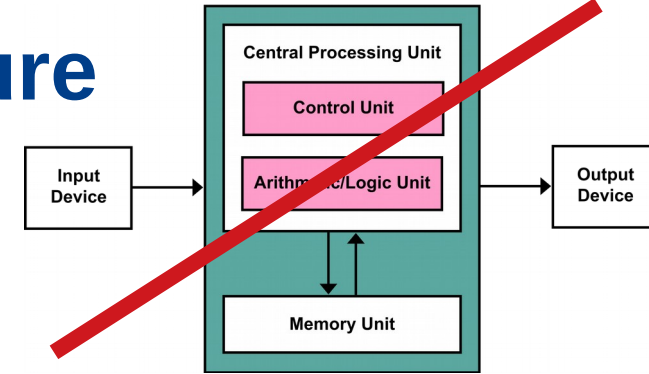
Beyond von-Neumann Architecture

Characteristics, e.g.:

- w/o sequential flow of control
- w/o the concept of a named storage variable

Programming examples (non-procedural):

- declarative (properties)
- data-driven (DSP, analog, quantum gates)



<https://www.encyclopedia.com/computing/dictionaries-thesauruses-pictures-and-press-releases/non-von-neumann-architecture>; Kapooht, CC-BY-4.0 SA, Wikimedia:Von_Neumann_Architecture.svg; <https://www.ibm.com/quantum-computing/>; <https://newsroom.intel.com/news/intels-pohoiki-beach-64-chip-neuromorphic-system-delivers-breakthrough-results-research-tests/>

Potential Routes for Engagement

- **Programming Models:** Need continued community engagement
 - Describe and publish our algorithms and codes
 - Re-design and adopt to industry trends
 - Propose, influence and refine with scientific use-cases
- **Algorithms:** how could a Poisson-solve, PIC-push, advection-diffusion, beam-transport, QED processes be modeled with “X”?
- **Leave comfort zones:** efforts across natural sciences & engineering
- **Adopt:** codes, languages, mental models, unexpected abstractions, ...

References

- **Future Technologies Group (OLCF)**
https://extremecomputingtraining.anl.gov/files/2019/07/ATPESC_2019_Dinner_Talk_2_7-30_Vetter-The_Coming_Age_of_Extreme_Heterogeneity.pdf
- **Supercomputing Conference Panels:** Beyond Von Neumann, Neuromorphic Systems and Architectures
- **PASC 2019 Conference Keynote:** Flexibly Scalable High Performance Architectures with Embedded Photonics (K. Bergman)
- **Intel oneAPI:** FPGA; **SPCL** (ETH Zuerich): FPGA High-Level Synthesis
DARPA: IDEA/POSH Universal Hardware Compiler
- **The Networking & Information Technology Research & Development Program (NITRD),** nitrd.gov
- **TOP500.org** hpcg-benchmark.org
- **Blogs & online publishing:** karlrupp.net plasma.ninja/blog
hpcwire.com thenextplatform.com

Backup

Today and Near-Term Programming Models

- **Dominating HPC / Industry Programming Models**
 - trend towards mixed-functional programming
 - functional
 - declarative over prescriptive
 - array-oriented/declarative: unclear future
 - C++-based: zero-overhead abstractions, standardization
 - C/Fortran: slow adoption, few compilers
 - Gap bridged with compiler directives
 - Some continue to be adopted into JIT (Python, Julia, ...)
 - scalability (latencies; duplication or broadcasts)
 - Emerging HW: low-level (VERILOG, Assembly, ...) or special-purpose DSLs

NITRD 2019

- **AMD:**

- “It is possible to **abstract** and simplify much of the complexity so that programmers can utilize simpler models of a system and increase their productivity. Useful **software abstractions** improve developer productivity and reduce execution risk by muting the cognitive noise produced by complexity.” **≠ traditional levels of abstraction**

- **“A New Golden Age for Computer Architecture”**

ACM Turing Award laureates John Hennessy and David Patterson:

- “The next decade will see a Cambrian explosion of novel computer architectures, meaning exciting times for computer architects in academia and industry.”

NITRD 2019

- NVIDIA:
 - “Looking forward, the practical realization of slowing technology scaling will likely require a range of approaches including: (1) architectures that incorporate increasingly specialized accelerator hardware; (2) packaging, signaling, and interconnect technologies that enable greater scaling at both the “node” and “system” level; (3) novel devices (e.g., carbon nanotube FETs) that can provide smaller digital devices at lower power; and (4) novel computing technologies such as analog, quantum, and neuromorphic that may require fundamental changes to algorithms.”

Semiconductor Manufacturing

Number of Semiconductor Manufacturers with a Cutting Edge Logic Fab										
SiITerra										
X-FAB										
Dongbu HiTek										
ADI	ADI									
Atmel	Atmel									
Rohm	Rohm									
Sanyo	Sanyo									
Mitsubishi	Mitsubishi									
ON	ON									
Hitachi	Hitachi									
Cypress	Cypress	Cypress								
Sony	Sony	Sony								
Infineon	Infineon	Infineon								
Sharp	Sharp	Sharp								
Freescall	Freescall	Freescall								
Renesas (NEC)	Renesas	Renesas	Renesas	Renesas						
Toshiba	Toshiba	Toshiba	Toshiba	Toshiba						
Fujitsu	Fujitsu	Fujitsu	Fujitsu	Fujitsu						
TI	TI	TI	TI	TI						
Panasonic	Panasonic	Panasonic	Panasonic	Panasonic	Panasonic					
STMicroelectronics	STM	STM	STM	STM	STM					
HLMC	HLMC		HLMC	HLMC	HLMC					
UMC	UMC	UMC	UMC	UMC	UMC		UMC			
IBM	IBM	IBM	IBM	IBM	IBM	IBM				
SMIC	SMIC	SMIC	SMIC	SMIC	SMIC		SMIC			
AMD	AMD	AMD	GlobalFoundries	GF	GF	GF	GF			
Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung
TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC
Intel	Intel	Intel	Intel	Intel	Intel	Intel	Intel	Intel	Intel	Intel
180 nm	130 nm	90 nm	65 nm	45 nm/40 nm	32 nm/28 nm	22 nm/20 nm	16 nm/14 nm	10 nm	7 nm	5 nm